主动学习的科技文献研究对象标引体系研究*

贺惠新　刘丽娟
(同方知网(北京)技术有限公司　北京 100192)

摘要:【目的】识别论文标题中的研究对象属性实例，试图利用少量标注样本，最大限度地提高研究对象识别的准确率。【方法】分析科技文献中研究对象的语法特征，利用少量样本基于条件随机场序列标注算法，对研究对象进行识别和抽取，并引入基于未标注数据的主动学习的迭代标引体系，提高研究对象识别的准确率。【结果】能够高效利用未标注数据，并最大限度地提高研究对象识别的准确率，标注准确率达到 78.3%。【局限】算法运行效率有待进一步优化。【结论】对科技文献中研究对象属性实例具有较好的识别效果，为进一步挖掘科技文献中的知识体系和结构打下基础。

关键词: 科技文献　研究对象　条件随机场　迭代标引体系　主动学习
分类号: TP393　G25

1　引　言

科学论文是科研工作者在某一学术课题上，关于新的科学研究成果或创新见解的文字体现，是由作者通过书面撰写，总结提炼研究工作的展现形式。科学论文一般包括不同的研究元素，如研究背景、研究对象、研究过程、研究方法、研究结论等。其中论文的研究对象指论文主要研究目标的核心主体，能高效清晰定位出对应文章的关注面，包括客观事物、理论、事件、过程、关系等属性实例。研究对象的提取能够将论文的主要研究目标以直观的形式展现出来，有助于读者快速掌握这一对象的相关信息，方便检索和对比相关研究的内容。

本文针对论文中的研究对象属性实例进行识别和提取，并提出基于主动学习的标引体系，使用少量已标注样本进行研究对象属性标注，并充分利用大量未标注数据，在节省人工标注成本的基础上，最大限度地提高研究对象提取的准确率，此体系可为论文知识结构的自动抽取和组织管理提供借鉴。

2　相关研究

论文研究对象的抽取属于属性抽取的研究范畴，属性抽取隶属于细粒度知识抽取的研究范畴。就抽取对象类型而言，属性抽取主要分为对实体的属性抽取，如人物[1-2]、物品[3-4]等，以及对概念的属性抽取[5-6]。而对概念的属性抽取又可分为通俗概念的属性抽取[7]和学术概念的属性抽取[5-8]。本文抽取目标是医学领域论文，待抽取对象包括学术概念的属性实例，也包括一系列医学命名实体[9]组成的目标对象，如疾病、药物、治疗方法等。因此，本文以属性抽取为基本思路，从学术概念属性抽取和命名实体识别两方面，对相关研究进行介绍。在领域文章中进行属性抽取时，采用的方法主要包括基于规则的方法、机器学习的方法以及两者相结合的方法。

2.1　基于规则的方法

采用手工或者自动构建的规则，识别关系与概念

通讯作者：刘丽娟，ORCID：0000-0001-9240-3998，E-mail：337145047@qq.com。
*本文系国家自然科学基金项目“群体性突发事件预警的超网络方法研究”(项目编号:71473034)的研究成果之一。

之间的语言模式并依此制定抽取规则。Fundel 等[10]制定候选关系的规则，从 Medline 摘要中抽取基因–蛋白质的关系，其利用基于规则的方法，并制定过滤方法对结果进行过滤，达到属性抽取的目的。但由于候选关系是由人工选定的，必定存在局限性，另外受现有分词工具效果的影响，准确率也会有所影响。张晗等[11]利用关联规则对医学数据进行知识抽取，分析词组配模式，利用文献资料抽取出 4 种肿瘤药物的主副主题词的语义关系搭配模式，利用这些关系模式达到属性关系抽取的目的。

2.2 机器学习的方法

将属性抽取问题转化为分类问题或标注问题，需选取特定特征并利用预先标注好的数据训练模型。CRFs 是一种概率图模型[12]，能够较好地表达元素之间的长距离依赖，从模型中抽取领域知识，避免最大熵隐马尔可夫模型(MEMM)和其他的条件马尔可夫模型会出现的标识偏置问题。孟洪宇等[13]以《伤寒论》为对象，采用条件随机场的术语自动识别方法，对特征进行对比实验，建立中医术语的自动识别模型。张帆等[14]借助领域本体或词表识别出的具有层级关系的主题词，识别创新点句中主题对应的属性实例，并采用一种语义标注、依存句法分析以及领域本体属性类相结合的方法，提高属性实例识别的准确率。

2.3 规则和机器学习相结合的方法

概念或命名实体属性抽取在其他学科中也有较广泛应用。在自然科学领域中，如计算机[15]、自然学[16]、分子材料学[17]等领域，通常属性抽取在本体构建、问答系统、自动摘要系统中起重要作用。Pham 等[15]利用基于涟波下降规则的方法建立文本标注规则，并利用正则表达式书写的规则进行过滤，通过分析不同标注类别的具体特征，对分类结果进行判定。Pechsiri 等[16-17]的研究对象是因果的属性关系，分析因果关系中的动词连接词语，对因果关系进行动词、原因、结果的标注，并利用贝叶斯分类器对动词连接的描述是否是原因和结果进行判断。Xiao 等[18]研究纳米材料对环境的影响，预先选定与纳米毒害性相关的 6 种实体以及 3 种属性，以段落为抽取对象，提取实体与属性之间的关系以及属性值。

在社会科学领域，部分学术概念属性较为抽象或具有主观性[8]，抽取前需确定属性的各类特征，如语言描述特征以及位置分布特征等。丁君军等[8]通过人工构建规则的方法，形成属性抽取的九大类描述规则，并针对《情报学报》的发表论文，进行学术文献中学术概念的抽取。程紫光[19]等利用 Bootstrapping 方法的和内模式的命名实体识别方法，针对特定领域实现半监督的命名实体识别。

机器学习的方法在训练数据的过程上，又可分为监督学习[20-21]、半监督学习[22-23]、无监督学习的训练方法。监督学习的方法是完全利用人工标注的训练数据，对模型参数进行估计，从而达到对未知数据的预测，不仅需要大量标注数据作训练集以保证泛化能力，同时标注也非常耗时耗力。特别是对于特定学术领域的标注，通常需要标注者有一定的背景知识。无监督学习无需人工标注数据，采用规则及其他方法抽取属性集合或关系集合，往往由于规则的限制导致效果不好。半监督的方法是采用两种方法的结合，充分发挥两种方法的优点，能够节省人工标注的成本，同时提高属性抽取的准确率。

借鉴关于属性抽取相关的成果，参考了机器学习，尤其是主动学习和 CRFs 序列标注的一些研究成果，初步探讨如何使用少量已标注样本进行研究对象属性标注；如何通过阈值估计，从大量未标注集合中选择有价值的样本进行人工标注，并尽可能节省人工成本，取得尽可能好的效果。通过构建 CRFs 的基于字的特征，训练模型，从而对论文研究对象属性进行抽取，并利用主动学习方法进行阈值估计，对数据集进行人工标记，以提高准确率。

3 研究对象生成标引体系

论文标题是最能简明扼要地反映论文中最重要的研究内容的逻辑组合，包括能够深刻揭示文章主要研究内容的关键词语，以及可以提供检索的特定实用信息。因此，论文标题是提取研究对象的主要目标。

针对论文标题中具有代表性意义的概念属性，对论文标题中的研究对象进行提取。通过分析论文标题中研究对象的语义特点和位置特征，对论文标题进行语义标注，同时采用主动学习的标注生成体系，使用少量已标注样本进行研究对象属性标注，充分利用大量未标注数据，最大限度提高生成标注的准确率，并与基于隐马尔可夫模型(HMM)的提取方法进行比较。

3.1　利用规则的策略进行研究对象提取

针对医学类中文学术论文，分析论文题目和研究对象，从论文题目中抽取子序列串组成的单个或多个连续字符作成研究对象。

基于规则的研究对象提取策略，采用的方法是利用规则的方法提取研究对象，由语料的标注结果发现，大部分研究对象都是从题目中提取医学类专业词语或者关系，去掉大部分常用词语，利用连接词、介词等切分而成的子串，所以考虑建立常用词语表，利用常用词语表，去除论文题目中的常用词语，并利用介词、连词、助词等无实际意义的词语，对论文题目进行切割，提取得到一个或多个研究对象。同时编写正则表达式对模型标引有误的具有明显句式结构的标题进行过滤。

此方法取得了一定的效果，使得利用介词、连词、助词等切分的一部分数据能够被正确分割开。其效果的局限性及原因如表 1 所示:

表 1　利用规则的策略的局限性

原因	解释
常用词的定义不明确	有些词语在一部分标题中属于常用词语，在另一部分中却不是常用词语。所以利用词典匹配的方法达到的效果也是有限的。
没有考虑词语的前后语义	利用介词、连词、助词等进行切割，由于没有考虑词语的前后语义，当两个连接词语同时修饰一个词语时，该方法并不有效。而且研究对象的提取由于句式的多样性，往往分为多种情况，单纯利用词语分割以及规则的方法也达不到理想的效果。

3.2　基于条件随机场的序列标注算法

条件随机场(Conditional Random Fields, CRFs)是由 Lafferty 等[12]于 2001 年提出的一种用来标注和划分序列结构数据的概率化结构模型，在自然语言处理领域得到了广泛的应用。

采用的基础标注算法为基于条件随机场[12]的序列标注算法，其中对准确率影响最大因素仍然是关于特征选取。实验的基本特征单位是“字”。根据实验为每个特征字添加了特征，如表 2 所示。

3.3　基于主动学习的研究对象生成标引体系

主动学习的过程为：在已经标好类标的数据集 K(初始时可能为空)和还没有标记的数据集 U 中，通过 K 集合的信息，找出一个 U 的子集 C，提出标注请求，待专家将数据集 C 标注完成后加入到 K 集合中，进行下一次迭代。

表 2　基于 CRFs 的序列标注算法特征

特征	描述
字类型特征	包括汉字、数字、字母、标点符号、大写的数字。以及字在标题中的相对位置的数值表示，大小为 0 到 1 之间。
词语特征	将字所在标题进行分词后，此字所归属的词的词性，及此字所在词语的位置信息：字在词语的开头、中间或者结尾。
词典词特征	训练语料中的高频词构成的词典中的词，在特征字所在句子中是否出现。
固定表述结构	与此字的距离(如“谈”，“轮”，“基于”等词)。
最后四字特征	此字所在标题句的最后 4 个字。
unigram 特征	此字位置为 0，对应包括(–2, –1, 0, 1, 2)等 5 个位置的字的词组。
bigram 特征	此字位置为 0，对应包括(–2*–1, –1*0, 0*1, 1*2)等 4 个组合的词组。

采用基于主动学习的方法进行实验，基于 CRFs 模型对已标注数据建模，进而对未标注数据进行预测，从大部分未标注集合中挑选尽量少的部分数据进行人工标注，并将标注后的结果再加入，进行迭代重新建模，最大限度地提高准确率。

在用每次训练生成的 CRFs 模型对新数据进行标注时，判别阶段是比较各字归属于不同类别的概率，并选取最大概率的类别为标注结果，而最大概率类别与次大概率类别的概率差值，本质上是可用于评判每次模型分类置信度的依据。本文选择要新标引的数据的方式即是基于分析此概率差值。

(1) 概率差值的分析

为了计算最优的阈值，将数据分成三组：训练数据 a，用作训练模型；添加数据 b，未知数据筛选，添加到训练集；测试数据 c，比较前后实验的准确率差值。通过对 a 建模对 c 预测，准确率为 P_1，并在 b 中选择 $\varepsilon \in [0, \partial_i]$ 的数据添加到 a 中，重新建模并对 c 预测，准确率为 P_2，观察两次准确率之差 $\Delta P = P_2 - P_1$，根据 ΔP 的大小差值判断该 ε 区间数据是否对模型有效。为了节省人工标注成本，须尽可能减少人工标注的成本，即筛选出进行人工标注的数据量 N 应尽可能少。引入如下判别公式:

$$R = \arg\max_{\varepsilon} \frac{\Delta P}{N} \tag{1}$$

其中，ε =标签的概率最大值−次大值，$\Delta P = P_2 - P_1 = f(\varepsilon)$ 表示前后两次模型准确率的差值，$N = g(\varepsilon)$ 为添加的人工标记的数据的数量。ε 为参数大小。当 ΔP 越大，N 越小时，越能添加尽量少的数据量，得到最好的实验效果。即当 R 值越大时，这时的 ε 取得最优值。

(2) 主动学习的迭代训练过程

迭代训练的过程为：通过阈值估计的 ε 值对未知数据进行筛选，并进行人工标注，添加到训练数据中重新进行模型参数估计。通过多次迭代该过程，使得在数据量和准确率之间达到平衡，即 R 值达到最大值。基于主动学习的研究对象生成标引体系的流程如图 1 所示：

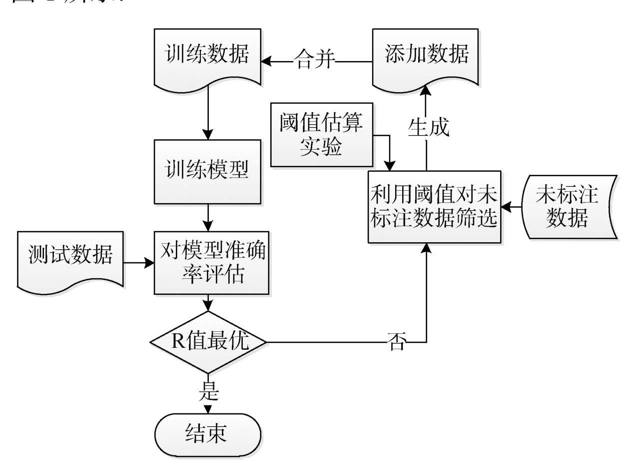


图 1 基于主动学习的研究对象生成标引体系流程

利用原始训练数据，进行模型参数估计，并利用测试数据对准确率进行评估，通过主动学习的迭代训练的方法，用已建立的模型对未知数据进行筛选，采用阈值估计实验得到阈值区间，挑选数据添加到训练集并重新进行参数估计，对新的模型进行测试。经过多次迭代，使得模型在准确率和训练效率上达到最优值。

4 实 验

4.1 实验数据

实验数据来源于中国知网[24]医学类学术论文，人工对论文标题进行研究对象的标注。选取 18 449 条作为最初的训练数据。

4.2 实验设置及结果分析

(1) 阈值估计实验

前后两次实验准确率差值如图 2 所示：

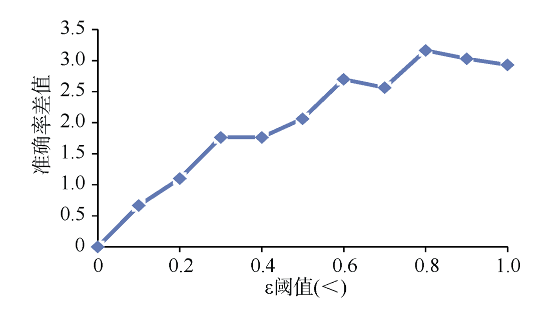


图 2 前后两次实验准确率差值

训练数据的变化如图 3 所示：

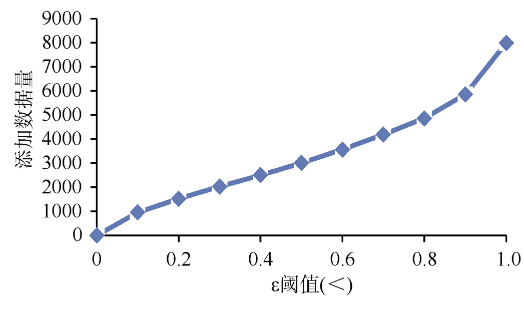


图 3 训练数据的变化

其中，R 的变化趋势如图 4 所示：

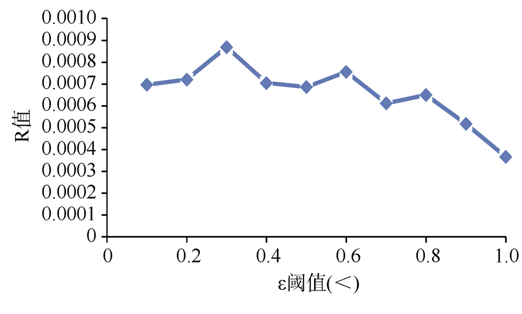


图 4 R 的变化趋势

通过实验结果可看出，准确率差值和添加数据量的变化规律。ε 在[0, 0.3]之间随着阈值的增大，添加标记的数据量越来越多，模型准确率呈上升趋势，表明这个区间数据对于模型的补充作用较大。当阈值在[0.3，0.8]时准确率的增速变缓；在[0.8，1]这段时间准确率差值波动起伏，表明该区间数据对于模型的干扰作用较大。

随着阈值的不断增大，数据量也是呈线性增长的趋势。通过计算得出 ε 在[0，0.3]时模型准确率提升了 1.7，而 ε 在[0.3，0.8]时准确率提升仅 1.3，且准确率波动明显，存在很多不确定性。数据量随着阈值的增大是呈线性增长的，当 ε 在[0，0.3]时添加的数据量和 ε

在[0.3, 0.8]时的比值约为 1∶3。实验目的是为了尽可能少添加人工标注的数量，降低人工成本，最大限度地提高模型准确率。可看出 ε 在[0, 0.3]时的数据对于模型的补充作用较大，而且需要人工标注的数据量也最小。由准确率变化趋势可看出 ε 在[0,0.3]区间模型准确率有大幅度提高，而要人工标注的数据也相对较少，故选择这个区间是合理的。

由图 4 可看出，R 的变化趋势是 ε 在[0, 0.3]区间呈上升趋势，当 ε>0.3 时，R 随之减小，故选择[0, 0.3]区间能尽量少地减少数据量，最大限度提高准确率。

(2) 主动学习的迭代训练实验及对比实验

单次实验中，实验采用基于初始实验数据进行五折交叉验证，在每份训练集上进行训练得到 CRFs 模型，并在对应的测试集上进行评判，最终计算平均准确率。

CRFs 算法窗口大小分别设置为 2, 4, 6, 8, 10, 12, 对比实验准确率。随着窗口的增大，预测准确率有一定程度提升，当窗口大小为 8-12 时，效果提高不明显，训练时间和内存占用却成倍增大，因此本实验采用最佳的窗口大小 6 进行实验。CRFs 算法采用 L-BFGS 参数估计算法，L1/L1 正规化系数 c1=0 和 c2=1。

迭代训练的过程为：通过上述阈值估计实验得出的 ε 阈值对[0, 0.3]区间的未知数据进行筛选，并进行人工标注，添加到训练数据中重新进行参数估计。通过多次迭代该过程，在数据量和准确率之间达到平衡，即 R 值达到最大值。基于本文提出的运用主动学习的迭代标引体系，对结果进行统计，准确率的变化趋势如图 5 所示：

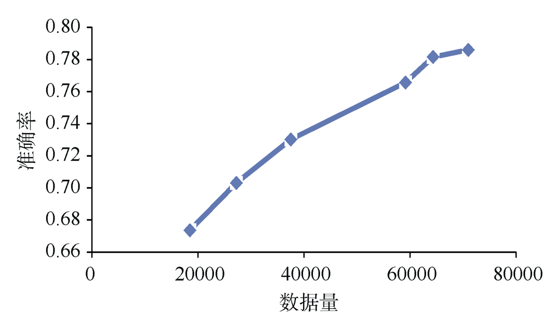


图 5 准确率变化趋势

在人工标注的数据基础上，初次模型提取研究对象的准确率为 67.5%，单纯采用 CRFs 标注方法得到的实验结果，初步达到一定的效果。随着迭代轮次的增加，准确率 P 在一定区间内呈上升趋势，说明随着主动学习方法添加了特定的数据，对原来模型的盲区产生针对性的补充，其训练时能获取的数据的特征空间也越来趋近于此特征的完备空间，从而使得数据的预测越来越准确。当数据量增加到一定程度，训练的特征信息越来越饱和，其增长的速率缓慢。而在 5 次迭代后，研究对象提取模型的准确率已达到 78.3%，极大提高了抽取的准确率。

与基于隐马尔可夫模型(HMM)的提取方法进行对比，如图 6 所示：

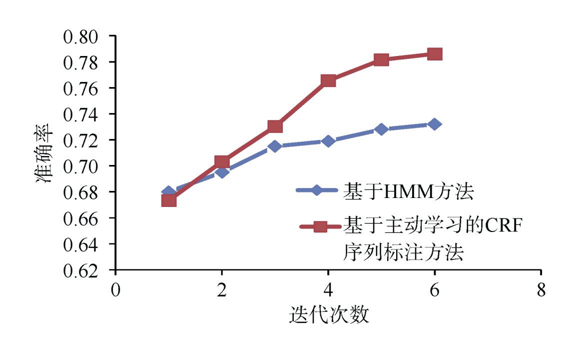


图 6 对比实验结果

基于主动学习的 CRFs 序列标注方法在每个数据段的准确率都高于基于隐马尔可夫模型的方法，整体性能也明显优于该方法。而且随着数据量的不断增加，基于主动学习的 CRFs 序列标注方法的准确率呈明显上升趋势。而 HMM 模型的假设前提在比较小的数据集上是合适的，但实际上在大量真实语料中观察序列更多是以一种多重的交互特征形式表现。由于实体本身结构所具有的复杂性，利用简单的特征函数往往无法涵盖所有的特性，导致其具有局限性。

结合图 5 和图 6 的实验结果，可得出该方法随着迭代次数的增加，准确率明显提升，并且通过与基于隐马尔可夫模型及单纯 CRFs 的方法进行对比，充分说明本文方法的有效性。

5 结 语

本文针对领域科技文献的元数据抽取问题，系统分析了研究对象的结构特点和语义特征，利用条件随机场的序列标注算法，提取出文献中的研究对象，同时提出一种主动学习的研究对象标引体系，从未知数据集中筛选有效数据，使模型达到最好的效果。本文方法不仅能够减少人工标注的成本，最大程度提高机

器学习算法的运行效率，同时能充分利用大量未标注数据，并使研究对象的提取得到最优的性能提升。本文方法不仅适用于医学领域科技文献，而且同样适用于其他领域科技文献的元数据抽取问题。而且主动学习的标引体系可用于指引其他元数据抽取问题，具有很强的借鉴意义。

参考文献：

[1] Lan M, Zhang Y Z, Lu Y, et al. Which Who are They? People Attribute Extraction and Disambiguation in Web Search Results [C]. In: Proceedings of the 18th World Wide Web Conference, Madrid, Spain. 2009.

[2] 李红亮. 基于规则的百科人物属性抽取算法的研究[D]. 成都：西南交通大学，2013. (Li Hongliang. Research on Character Attributes Extraction Based on Rules from Baidu Encyclopedia [D]. Chengdu: Southwest Jiaotong University, 2013.)

[3] 曾道建，来斯惟，张元哲，等. 面向非结构化文本的开放式实体属性抽取[J]. 江西师范大学学报：自然科学版，2013，37(3): 279-283. (Zeng Daojian, Lai Siwei, Zhang Yuanzhe, et al. Open Entity Attribute-Value Extraction from Unstructured Text [J]. Journal of Jiangxi Normal University: Natural Science Edition, 2013, 37(3): 279-283.)

[4] Ghani R, Probst K, Liu Y, et al. Text Mining for Product Attribute Extraction [J]. ACM SIGKDD Explorations Newsletter, 2006, 8(1): 41-48.

[5] 贾真，杨宇飞，何大可，等. 面向中文网络百科的属性和属性值抽取[J]. 北京大学学报：自然科学版，2014，50(1): 41-47. (Jia Zhen, Yang Yufei, He Dake, et al. Attribute and Attribute Value Extracted from Chinese Online Encyclopedia [J]. Acta Scientiarum Naturalium University Pekinensis, 2014, 50(1): 41-47.)

[6] 刘丽佳，郭剑毅，周兰江，等. 基于 LM 算法的领域概念实体属性关系抽取[J]. 中文信息学报，2014，28(6): 216-222. (Liu Lijia, Guo Jianyi, Zhou Lanjiang, et al. Domain Concepts Entity Attribute Relation Extraction Based on LM Algorithm [J]. Journal of Chinese Information Processing, 2014, 28(6): 216-222.)

[7] 丁玉飞，王曰芬，刘卫江. 面向半结构化文本的知识抽取研究[J]. 情报理论与实践，2015，38(3): 101-106. (Ding Yufei, Wang Yuefen, Liu Weijiang. Research on Knowledge Extraction for Semi-structure Text [J]. Information Studies: Theory & Application, 2015, 38(3): 101-106.)

[8] 丁君军，郑彦宁，化柏林. 基于规则的学术概念属性抽取[J]. 情报理论与实践，2011，34(12): 10-14. (Ding Junjun, Zheng Yanning, Hua Bolin. Academic Concept Attribute Extraction Based on the Rules [J]. Information Studies: Theory & Application, 2011, 34(12): 10-14.)

[9] Rebholz-Schuhmann D. Biomedical Named Entity Recognition, Whatizit [A]. // Encyclopedia of Systems Biology [M]. Springer New York, 2013: 132-134.

[10] Fundel K, Küffner R, Zimmer R. RelEx—Relation Extraction Using Dependency Parse Trees [J]. Bioinformatics, 2007, 23(3): 365-371.

[11] 张晗，路振宇，崔雷. 利用关联规则对医学文本数据库进行知识抽取的尝试——以四种抗肿瘤药为例[J]. 现代图书情报技术，2006(9): 49-52. (Zhang Han, Lu Zhenyu, Cui Lei. Knowledge Extraction from Medical Literature Database Using Association Rule Mining——Taking Four Anti-neoplastic Medicines as an Example [J]. New Technology of Library and Information Service, 2006(9): 49-52.)

[12] Lafferty J D, McCallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]. In: Proceedings of the 18th International Conference on Machine Learning. 2001.

[13] 孟洪宇，谢晴宇，常虹，等. 基于条件随机场的《伤寒论》中医术语自动识别[J]. 北京中医药大学学报，2015，38(9): 587-590. (Meng Hongyu, Xie Qingyu, Chang Hong, et al. Automatic Identification of TCM Terminology in Shanghan Lun Based on Conditional Random Field [J]. Journal of Beijing University of Chinese Medicine, 2015, 38(9): 587-590.)

[14] 张帆，乐小虬. 领域科技文献创新点句中主题属性实例识别方法研究[J]. 现代图书情报技术，2015(5): 15-23. (Zhang Fan, Le Xiaoqiu. Research on Recognition of Concept Attribute Instances in Innovation Sentences of Scientific Research Paper [J]. New Technology of Library and Information Service, 2015(5): 15-23.)

[15] Pham S B, Hoffmann A. Extracting Positive Attributions from Scientific Papers[A]. // Discovery Science [M]. Springer Berlin Heidelberg, 2004: 169-182.

[16] Pechsiri C, Kawtrakul A. Mining Causality for Explanation Knowledge from Text [J]. Journal of Computer Science and Technology, 2007, 22(6): 877-889.

[17] Pechsiri C, Piriyakul R. Explanation Knowledge Graph Construction Through Causality Extraction from Texts [J]. Journal of Computer Science and Technology, 2010, 25(5): 1055-1070.

[18] Xiao L, Tang K, Liu X, et al. Information Extraction from Nanotoxicity Related Publications [C]. In: Proceedings of the

2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China. 2013: 25-30.

[19] 程紫光. 面向领域知识库构建的实体识别及关系抽取技术[D]. 哈尔滨: 哈尔滨工业大学, 2014. (Cheng Ziguang. Research on Named Entity Recognition and Relation Extraction Facing to Domain-Oriented Knowledge Base Construction [D]. Harbin: Harbin Institute of Technology, 2014.)

[20] Xiao J, Su J, Zhou G D, et al. Protein-Protein Interaction Extraction: A Supervised Learning Approach [C]. In: Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine. 2005: 51-59.

[21] 张益嘉. 生物医学领域的信息抽取与复合物识别研究[D]. 大连: 大连理工大学, 2014. (Zhang Yijia. Information Extraction in Biomedical Literature and Protein Complex Identification [D]. Dalian: Dalian University of Technology, 2014.)

[22] Li Y P, Hu X H, Lin H F, et al. Learning an Enriched Representation from Unlabeled Data for Protein-Protein Interaction Extraction [J]. BMC Bioinformatics, 2010, 11(S2): 7-10.

[23] 闫紫飞, 姬东鸿. 基于 CRF 和半监督学习的中文时间信息抽取[J]. 计算机工程与设计, 2015, 36(6): 1642-1646. (Yan Zifei, Ji Donghong. Exploration of Chinese Temporal Information Extraction Based on CRF and Semi-supervised Learning [J]. Computer Engineering and Design, 2015, 36(6): 1642-1646.)

[24] 中国知网[OL]. [2015-06-25]. http://www.cnki.net/. (CNKI [OL]. [2015-06-25]. http://www.cnki.net/.)

作者贡献声明:

贺惠新: 提出研究思路, 设计研究方案, 论文最终版本修订;
刘丽娟: 采集、清洗、分析、处理数据, 进行实验, 起草论文。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据[1-3]见期刊网络版 http://www.infotech.ac.cn; 支撑数据[4]由作者自存储, E-mail：huixinhe@qq.com。

[1] 贺惠新, 刘丽娟. tools_url.txt. 研究工具 crfsuite 和 python-crfsuite 的网页链接.

[2] 贺惠新, 刘丽娟. title.xlsx. 实验论文标题原始数据.

[3] 贺惠新, 刘丽娟. X.txt. 实验论文标题转换为特征后对应的文件.

[4] 贺惠新. studyObject.xlsx. 实验论文标注的标题研究对象数据.

收稿日期: 2015-10-13
收修改稿日期: 2015-12-22

A Scientific Research Object Labeling System Based on Active Learning

He Huixin　Liu Lijuan
(Tongfang Knowledge Network Technology Co., Ltd. (Beijing), Beijing 100192, China)

Abstract: [**Objective**] This study aims to identify the research object attribute instance from the paper titles. With the help of limited labeled samples, we could maximumize the accuracy of research object recognition. [**Methods**] We first analyzed the grammatical features of scientific research objects based on conditional random field sequence labeling algorithm. Second, we recognized and extracted research objects using a small amount of samples. Finally, we introduced an active learning iterative labeling system based on unlabeled data to improve the research object recognition accuracy. [**Results**] The results showed that the proposed method could efficiently use the unlabeled data, and increase the accuracy of the research object recognition to 78.3%. [**Limitations**] The proposed algorithm needs to be further optimized to improve its efficiency. [**Conclusions**] The proposed method performed well on the research object attributes identification, which is the foundation for further mining the knowledge system and the structure of science and technology literature.

Keywords: Scientific literature　Research objects　Conditional Random Fields　Iterative labeling system　Active learning